ML at CEPC

李刚

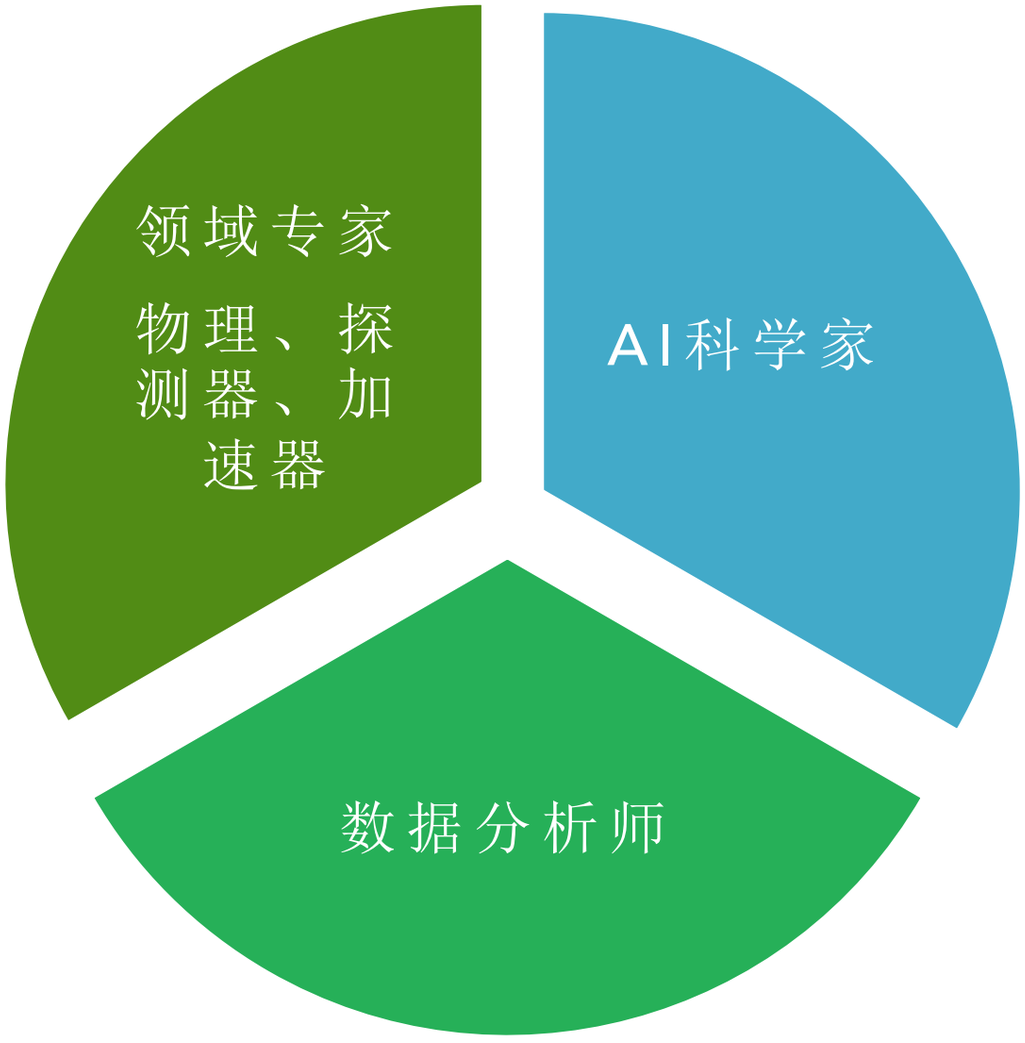
机器学习技术在高能所各学科中的应用研讨会

2022年09月18-19日

Outline

- Introduction
- ML at CEPC
- Summary

AI 相关领域的关系

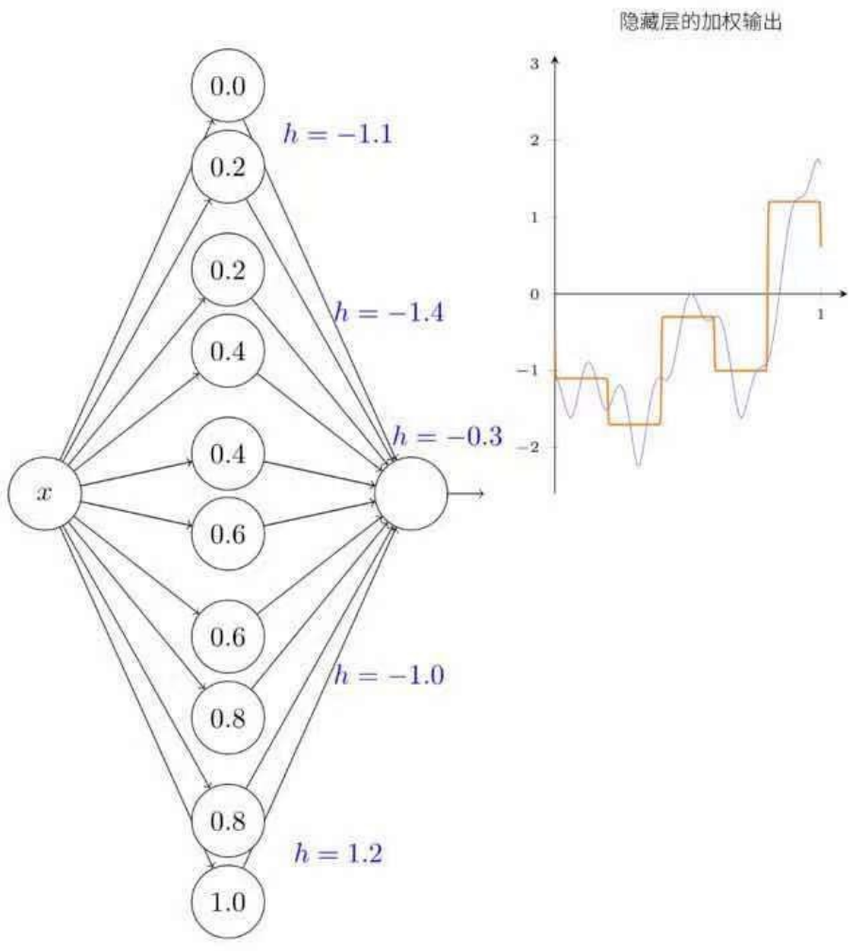


劣势：所有事情自己干

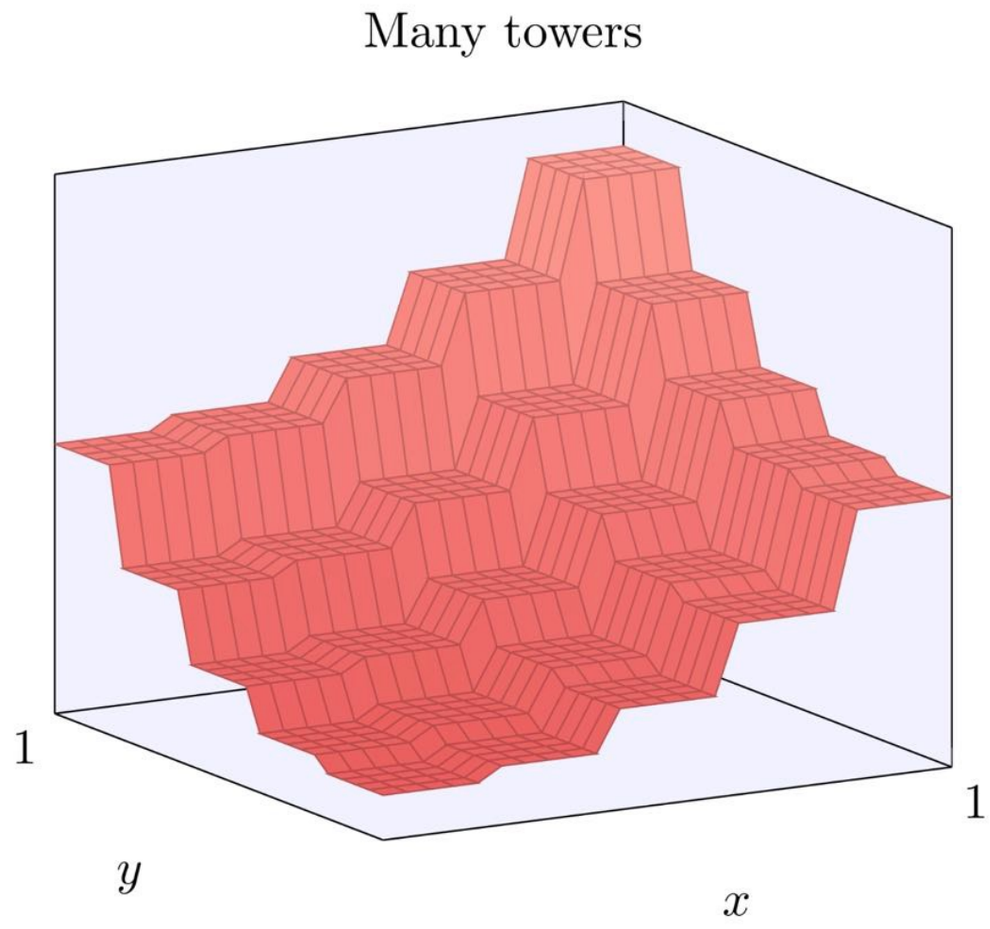
优势：自己干所有事情

Why works?

神经网络：万能函数展开器
Universal function approximator



一维



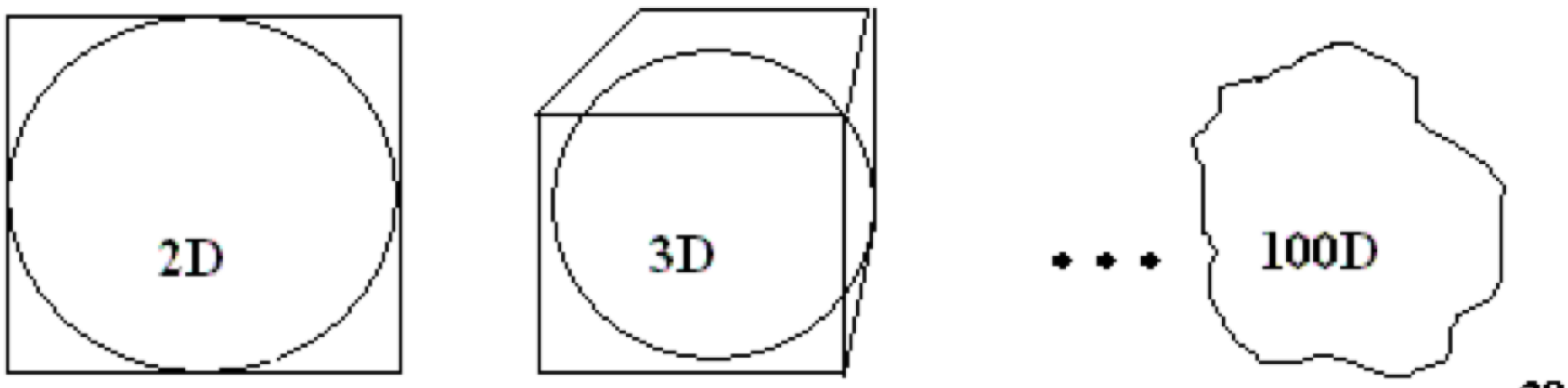
二维

可以以任意精度近似一个连续函数，如果隐藏层的神经元足够多的话 …

Why difficult?

Curse of dimensionality

Machine learning is about solving some standard mathematical problems, but typically in very high dimension!

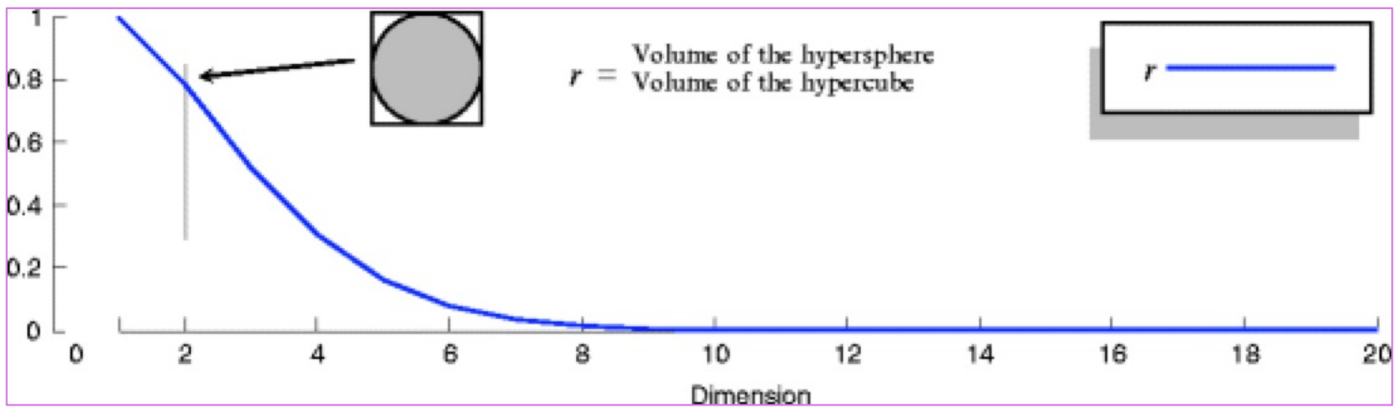


ratio: $4/\pi = 1.27$ ratio: $6/\pi = 1.91$ ratio: $4.2 \cdot 10^{39}$

$$\frac{A_{circle}}{A_{square}} = \frac{\pi}{4} \text{ for } d = 2$$

$$\frac{V_{sphere}}{V_{cube}} = \frac{\pi}{6} \text{ for } d = 3$$

$$\frac{V_{hypersphere}}{V_{hypercube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ as } d \to \infty$$



- D=1，100个平均分布的点能把一个单位区间以距离不超过0.01采样；
- D=10，则需要10^{20} 个采样点才能达到同样的采样率。

高维空间非常空旷，
“没有中心”，“都是角落”！

必须降维，
且大多具体问题可以降维！

What can be done?

模型没有好坏，但对具体问题有“偏好”

No Free Lunch Theorem

What does dinner cost? David Wolpert

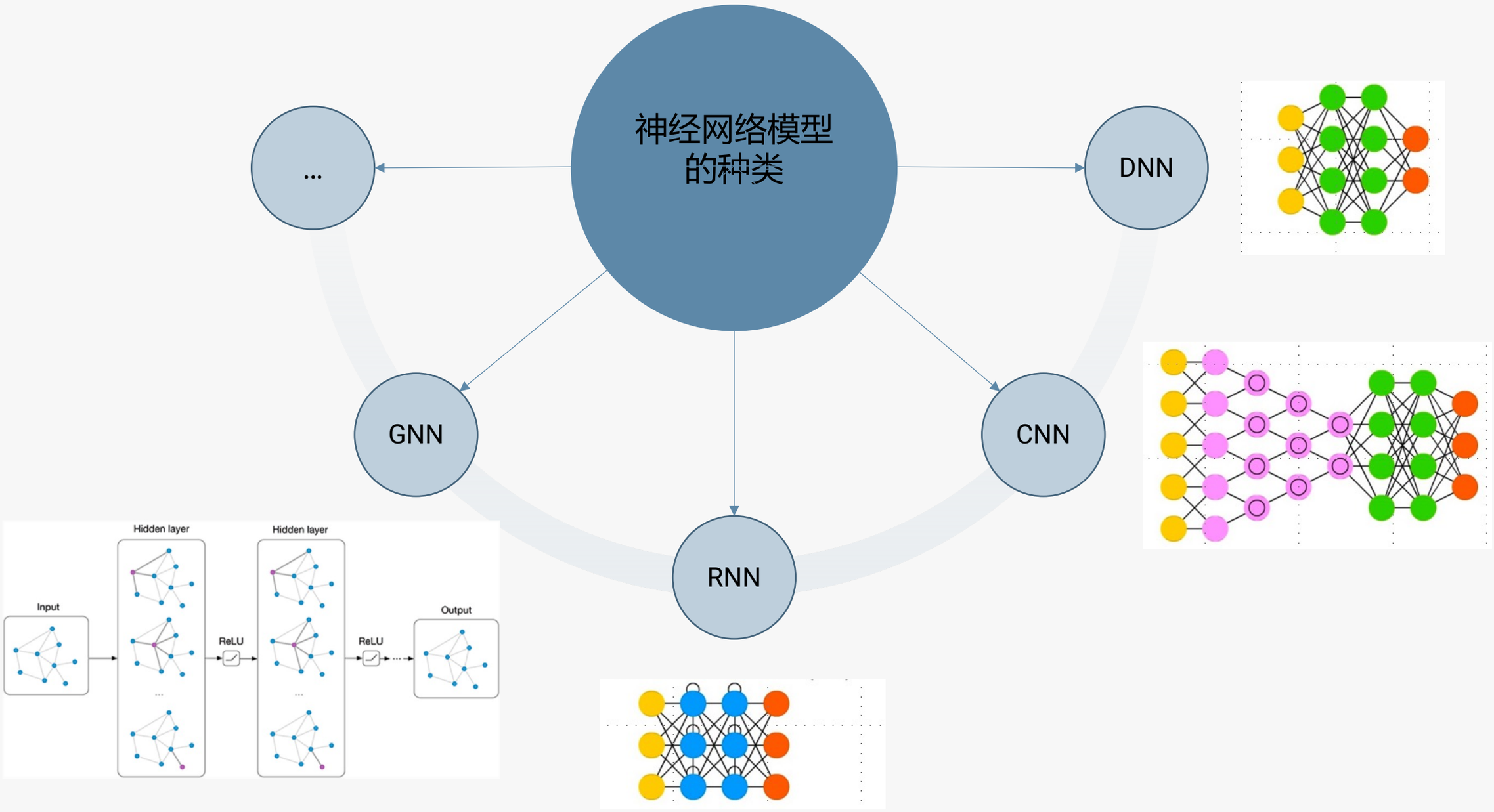
- 这是个严肃的数学定理
- 可以严格证明（假设：数据集是全集）

$$\sum_f P(d_m^Y \mid f, m, a) = \sum_f P(d_m^Y \mid f, m, a')$$

$$\forall\ a, a', d_m$$

- 但具体问题的数据一个子集
- 关键：合适的算法

用合适的算法



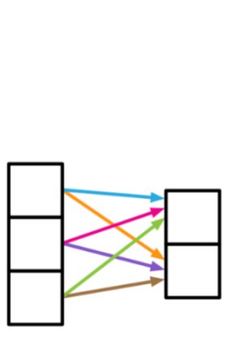
Relational inductive bias, deep learning, and graph networks
arXiv:1806.01261

用合适的算法

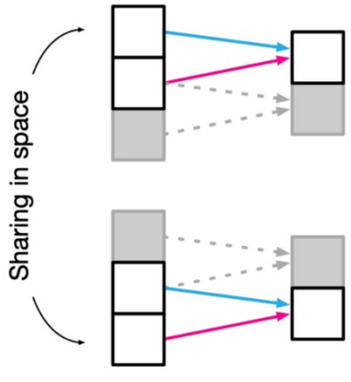
归纳偏好
Inductive bias

- ✓ NFL定理指出学习是不可能的，除非有先验知识。
- ✓ 通常情况下我们不知道上帝函数，但猜测它属于一个较小的假设类别。
- ✓ 这种基于先验知识对目标模型的判断就是 inductive bias —— 归纳偏好。归纳偏好所做的事情就是将无限可能的目标函数约束在一个有限的假设类别当中。
- ✓ 如果给出更加宽松的模型假设，也即用更弱的 inductive bias，那我们更有可能得到强力模型 —— 更接近目标函数 F，但是训练变得非常困难，乃至不可能。
- ✓ 学习者的归纳偏好是一组额外的假设，足以证明其归纳推理是演绎的推论。

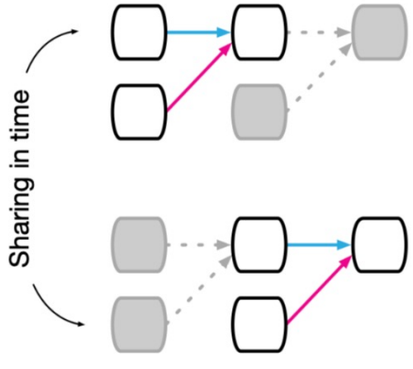
不同算法的归纳偏好比较



(a) Fully connected

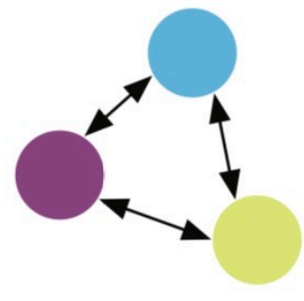


(b) Convolutional



(c) Recurrent

三体系统的质心



类型	输入形式	关系	偏好	变换	评论
DNN	单位	All-to-all	弱	---	信息无重用，无孤立
CNN	均匀像素	局域	取决局域性	空间平移	局域特征和平移不变性
RNN	均匀时间步长	序列	取决于序列性	时间平移	信息重用+时间平移不变
GNN	节点	边	变化大	点、边的交换	顺序无关，IB来自某种东西的“无”（absence）

What can be done?

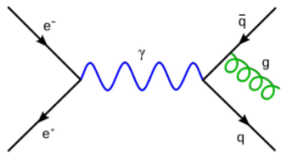
机器学习是统计学习，有其数学基础，很强大，也有缺陷，不是魔法盒子。统计、尤其是高维统计很有帮助。

机器学习模型对不同问题有的合适，有的不合适。需要理解自己科学问题，进行适当的建模，然后寻找市场上合适的模型或者自己设计模型。

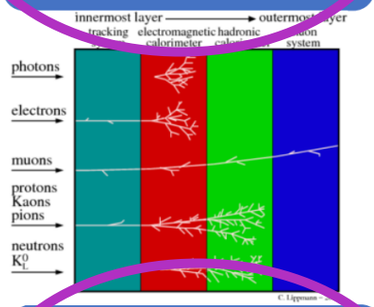
我们所能做的是：在既有框架下寻找和使用对当前问题有利的 inductive bias。

ML@HEP

nature

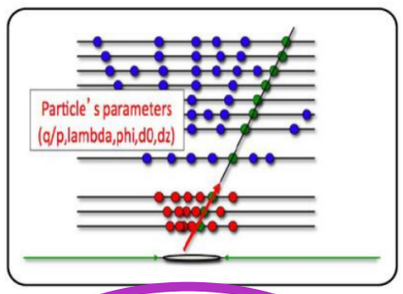


detector



event generator

simulation of detector response



reconstruction

hits→ final state particle :
Cluster hits
Classify clusters as particles
Infer/regress properties

Classify signal/background

event selection

Statistical inference

Hypothesis testing
Parameter estimation

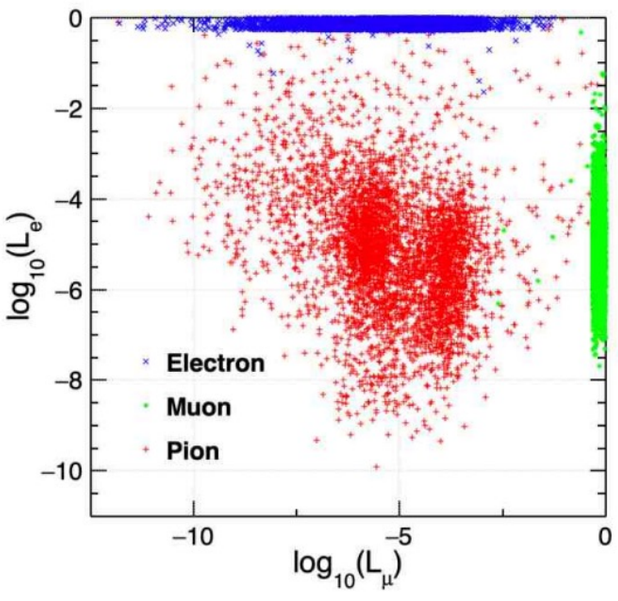
ML@CEPC

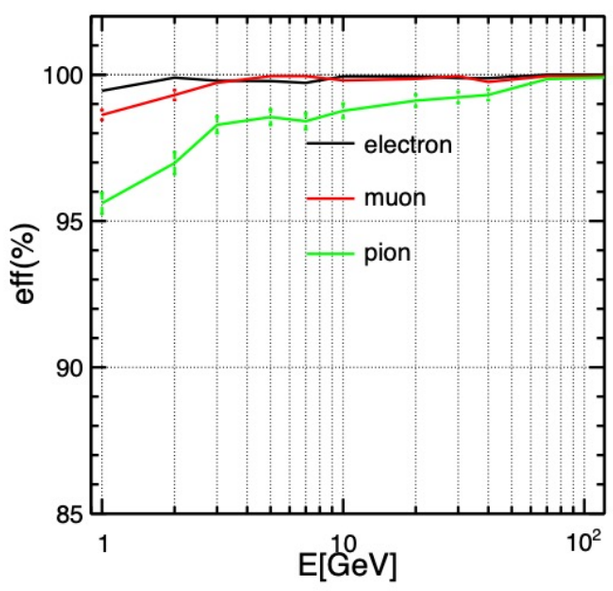
- Classification
  - PID
  - Jet flavor tagging
  - Event classification
- Pattern recognition
  - Using RNN to reconstruct peaks of primary ionization
- Background suppression + data compression
- Simulation

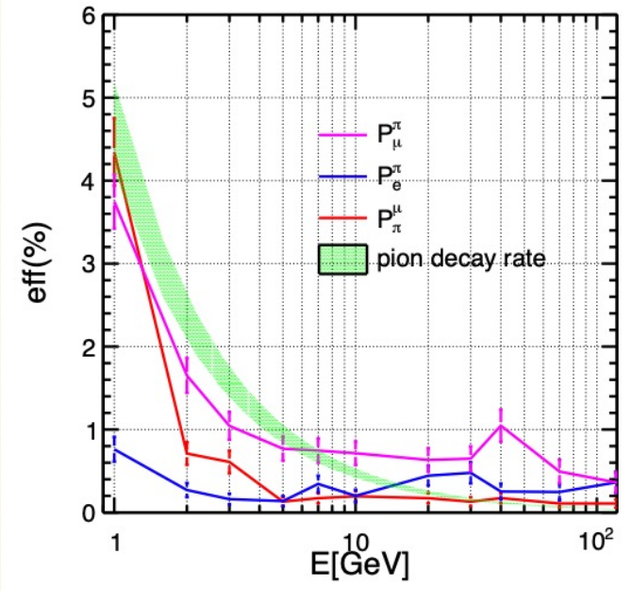
粒子分类

- TMVA + hand engineering features

- LICH uses TMVA methods to summarize 24 input variables into two likelihoods, corresponding to electrons and muons.
- The efficiency for electron and muon is higher than 99.5% (E>2 GeV). Pion efficiency ~ 98%.







Migration Matrix at 40GeV (LICH)

Type	e^- *like*	μ^- *like*	π^+ *like*
e^-	99.71 ± 0.08	< 0.07	0.21 ± 0.07
μ^-	< 0.07	99.87 ± 0.08	0.05 ± 0.05
π^+	0.14 ± 0.05	0.35 ± 0.08	99.26 ± 0.12

Migration Matrix for ALEPH PID ($> 2GeV$)(*Eur.Phys.J.C20:401-430,2001*)

Type	e^- *like*	μ^- *like*	π^+ *like*	undefined
e^-	99.57 ± 0.07	< 0.01	0.32 ± 0.0	0.09 ± 0.04
μ^-	< 0.01	99.11 ± 0.08	0.88 ± 0.08	0.01 ± 0.01
π^+	0.71 ± 0.04	0.72 ± 0.04	98.45 ± 0.06	0.12 ± 0.03

Dan Yu

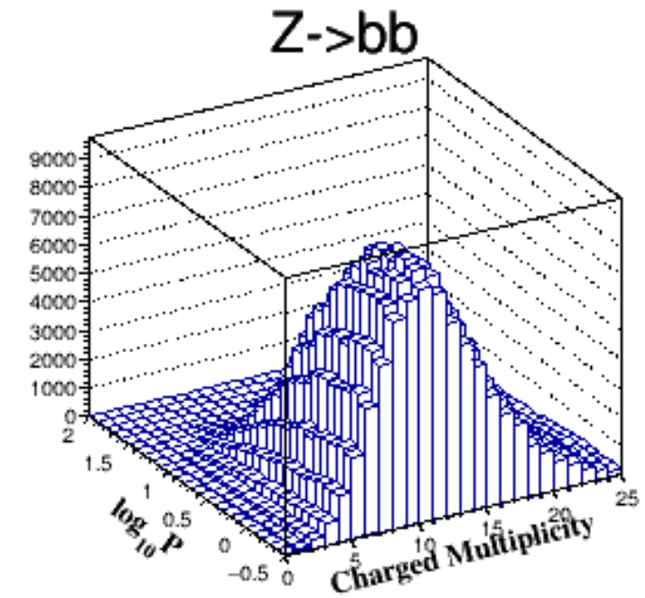
Jet 分类

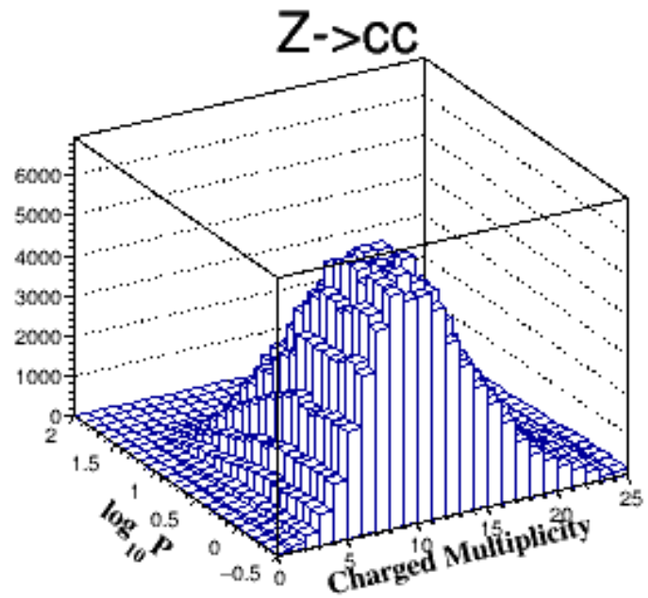
ArXiv:2208.13503, submitted to EPJC

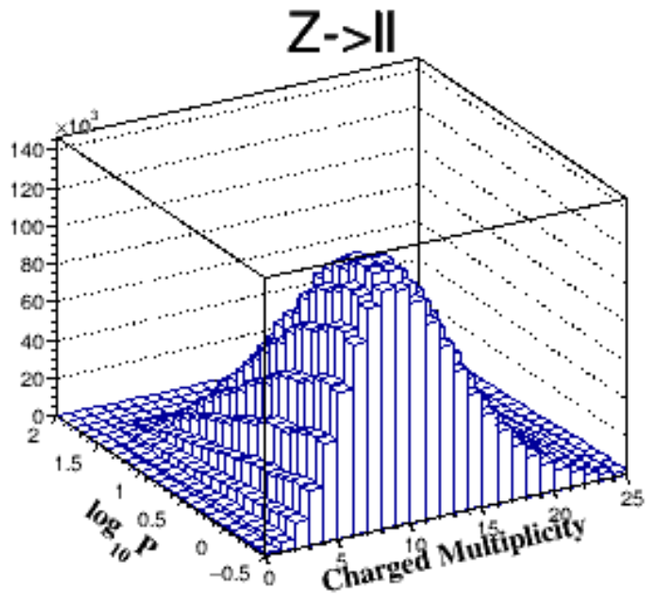
数据集

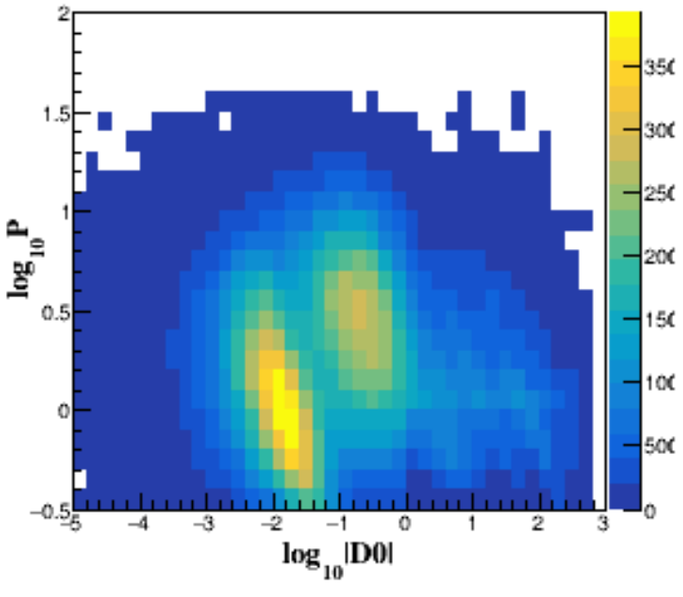
- 91 GeV
- Z → bb, cc, oo (uu,dd,ss)
- WHIZARD 产生/全模拟/重建
- Jet Clustering
- 每种样本 450k 事例（900k jets）

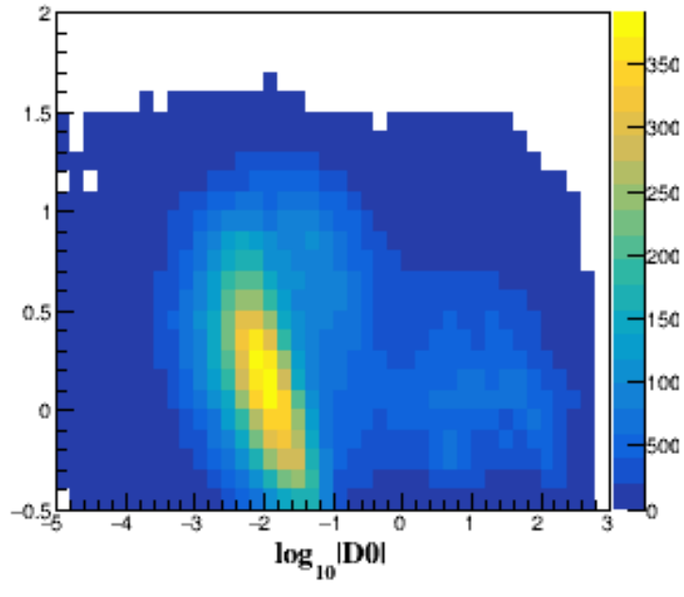
数据集中的 features

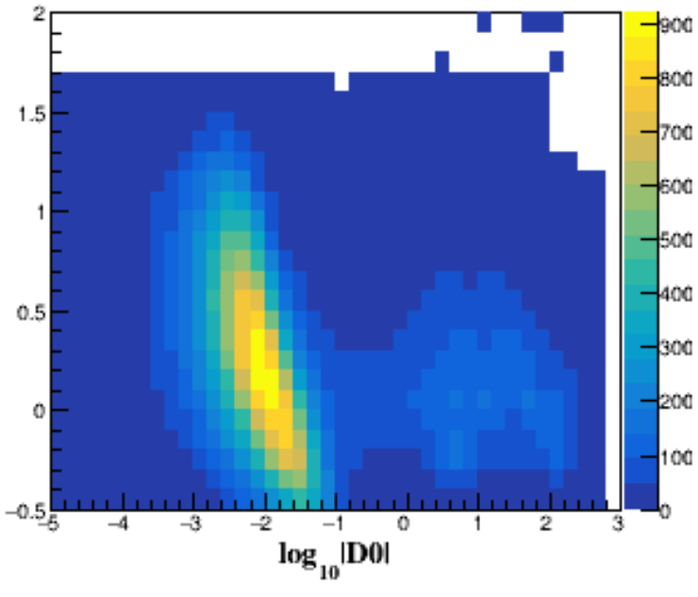




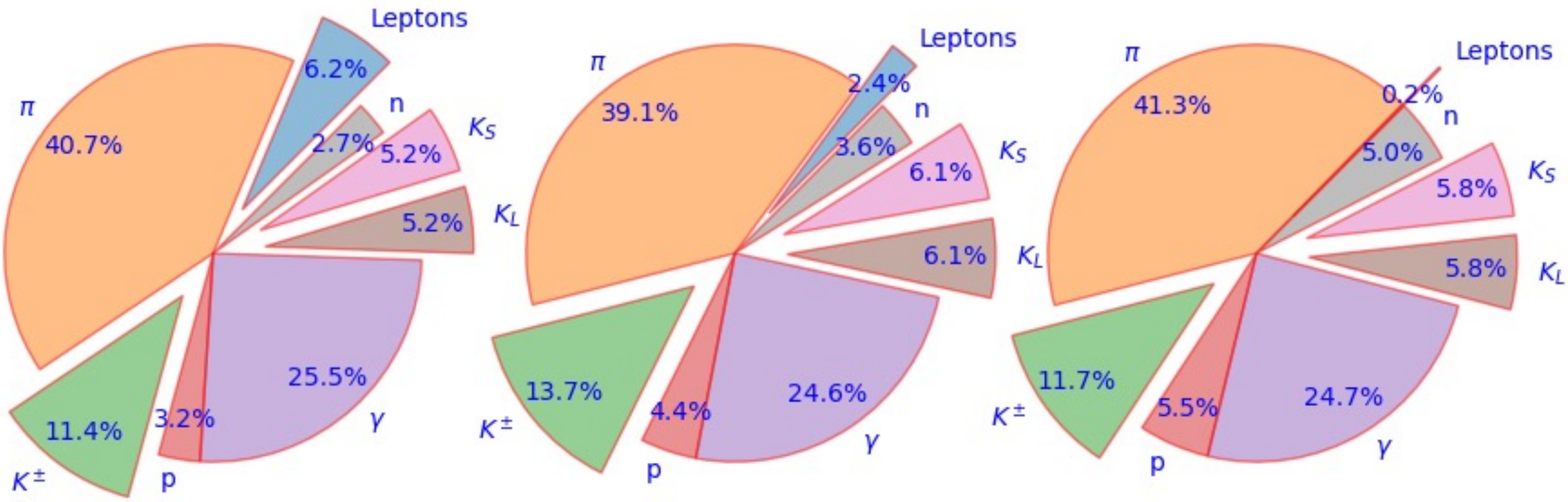








数据集中的 features



粒子种类特征

能看到的非常有限，更多的还需要算法去挖掘

不同算法结果比较（一）

Algorithm	ParticleNet	PFN	DNN	BDT	GBDT	gcforest	XGBoost
Accuracy	0.872	0.850	0.788	0.776	0.794	0.785	0.801

>0.90 @ fast sim

不同算法结果比较（二）

tag	$\epsilon_S(\%)$	$\epsilon \times \rho$			
		LCFIPlus	XGBoost	ParticleNet	PFN
b	60	-	-	0.589	0.596
	70	-	-	0.694	0.689
	80	-	0.747	0.780	0.763
	90	0.72	0.713	0.810	0.752
	95	-	0.609	0.721	0.645
c	60	0.36	-	0.548	0.485
	70	-	-	0.589	0.497
	80	-	0.345	0.584	0.467
	90	-	0.292	0.516	0.402
	95	-	0.251	0.451	0.348

简单估算c-tag：

sqrt(0.584/0.345)=1.3

统计误差减小 30%

$$\frac{1}{(\Delta\sigma_s)^2} = \frac{1}{\sigma_s}\mathcal{L}\epsilon_s\rho = \frac{1}{\sigma_s^2}S_{\text{tot}}\epsilon_s\rho$$

为什么效果好？——对称性

无序、数目可变的粒子集合

交换 对称性

$$J\left(\{p_1^\mu, \ldots, p_M^\mu\}\right) = J(\{p_{\pi(1)}^\mu, \ldots, p_{\pi(M)}^\mu\})$$

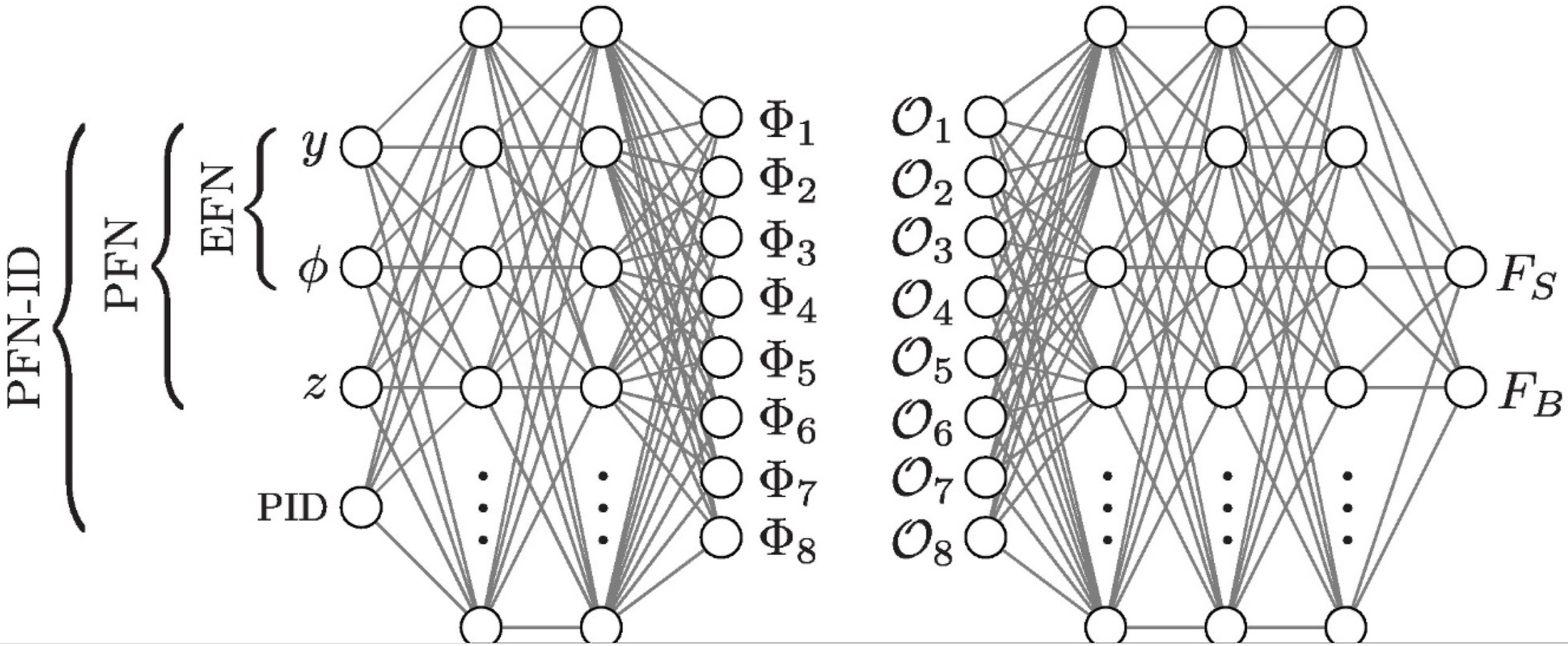
p_i^μ **represents *all* the particle properties:**

四动量，粒子质量，电荷，顶点信息（~10 个 float ）

为什么效果好？——信息多

Mapping

DNN



表示学习

分类

希格斯事例分类

arXiv:2105.14997， accepted by CPC

Higgs： 4production modes，9 decays modes

Prod\decay	cc	bb	μμ	ττ	γγ	gg	WW	ZZ	γZ
eeH									
μμH									
ττH									
qqH									

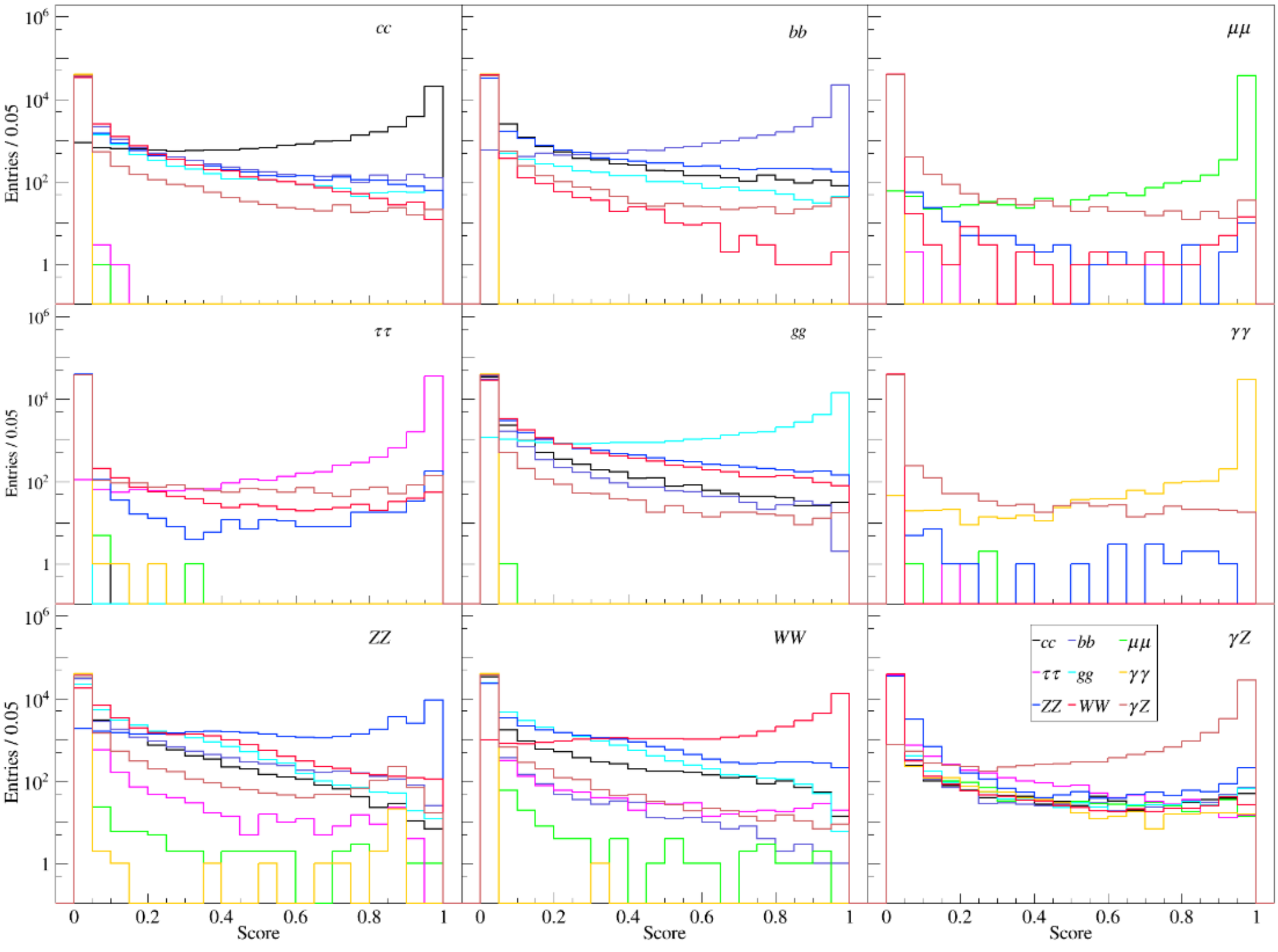
从 eeH开始尝试 9 分类

基于 Graph Neural Networks

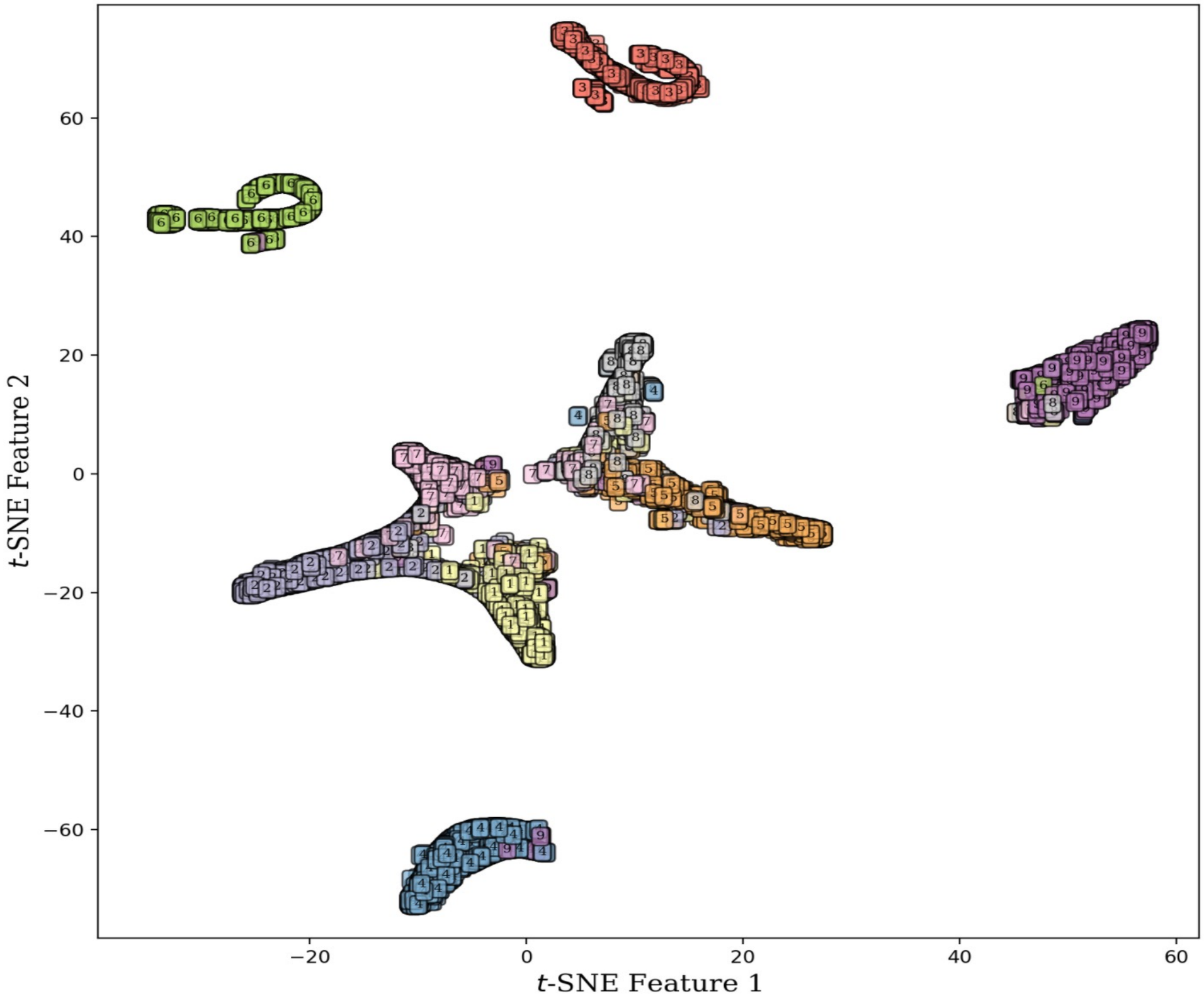
ParticleNet, PFN

希格斯的9个衰变模式，快速模拟

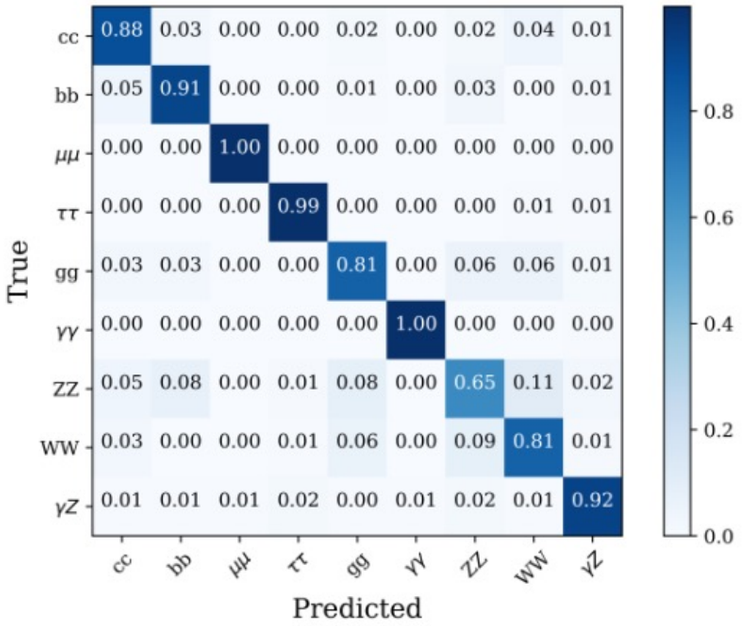
(eeH)



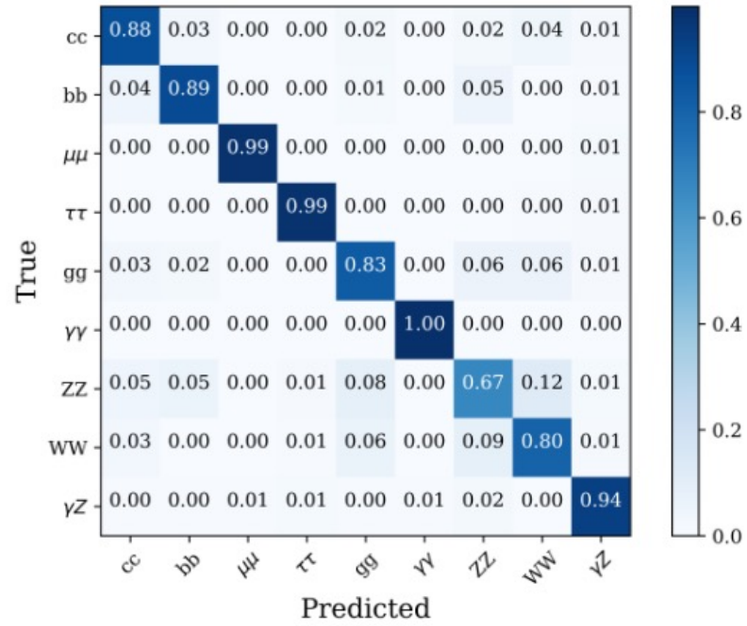
降维后的结果



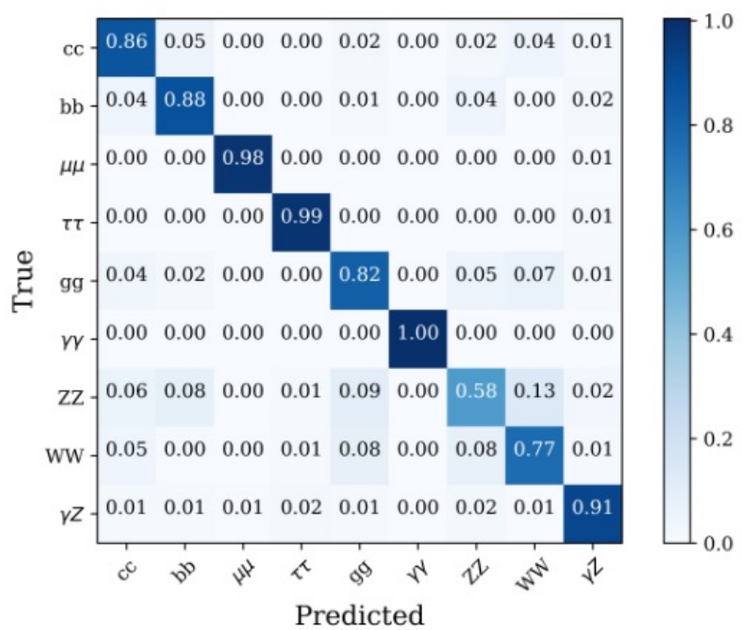
4个9分类


cc 0.88 0.03 0.00 0.00 0.02 0.00 0.02 0.04 0.01
bb 0.05 0.91 0.00 0.00 0.01 0.00 0.03 0.00 0.01
μμ 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00
ττ 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.01 0.01
gg 0.03 0.03 0.00 0.00 0.81 0.00 0.06 0.06 0.01
γγ 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
ZZ 0.05 0.08 0.00 0.01 0.08 0.00 0.65 0.11 0.02
WW 0.03 0.00 0.00 0.01 0.06 0.00 0.09 0.81 0.01
γZ 0.01 0.01 0.01 0.02 0.00 0.01 0.02 0.01 0.92
True
cc bb μμ ττ gg γγ ZZ WW γZ
Predicted
0.0 0.2 0.4 0.6 0.8

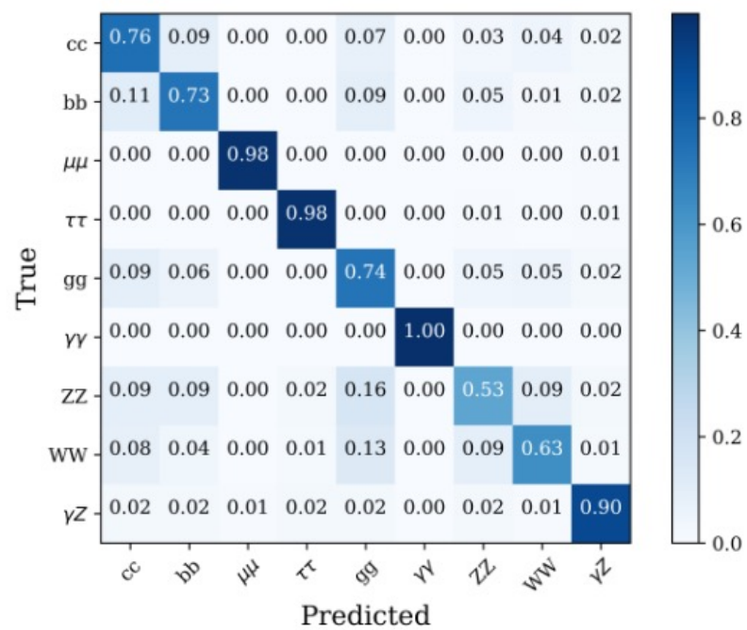
(a)


cc 0.88 0.03 0.00 0.00 0.02 0.00 0.02 0.04 0.01
bb 0.04 0.89 0.00 0.00 0.01 0.00 0.05 0.00 0.01
μμ 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.00 0.01
ττ 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.01
gg 0.03 0.02 0.00 0.00 0.83 0.00 0.06 0.06 0.01
γγ 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
ZZ 0.05 0.05 0.00 0.01 0.08 0.00 0.67 0.12 0.01
WW 0.03 0.00 0.00 0.01 0.06 0.00 0.09 0.80 0.01
γZ 0.00 0.00 0.01 0.01 0.00 0.01 0.02 0.00 0.94
True
cc bb μμ ττ gg γγ ZZ WW γZ
Predicted
0.0 0.2 0.4 0.6 0.8

(b)

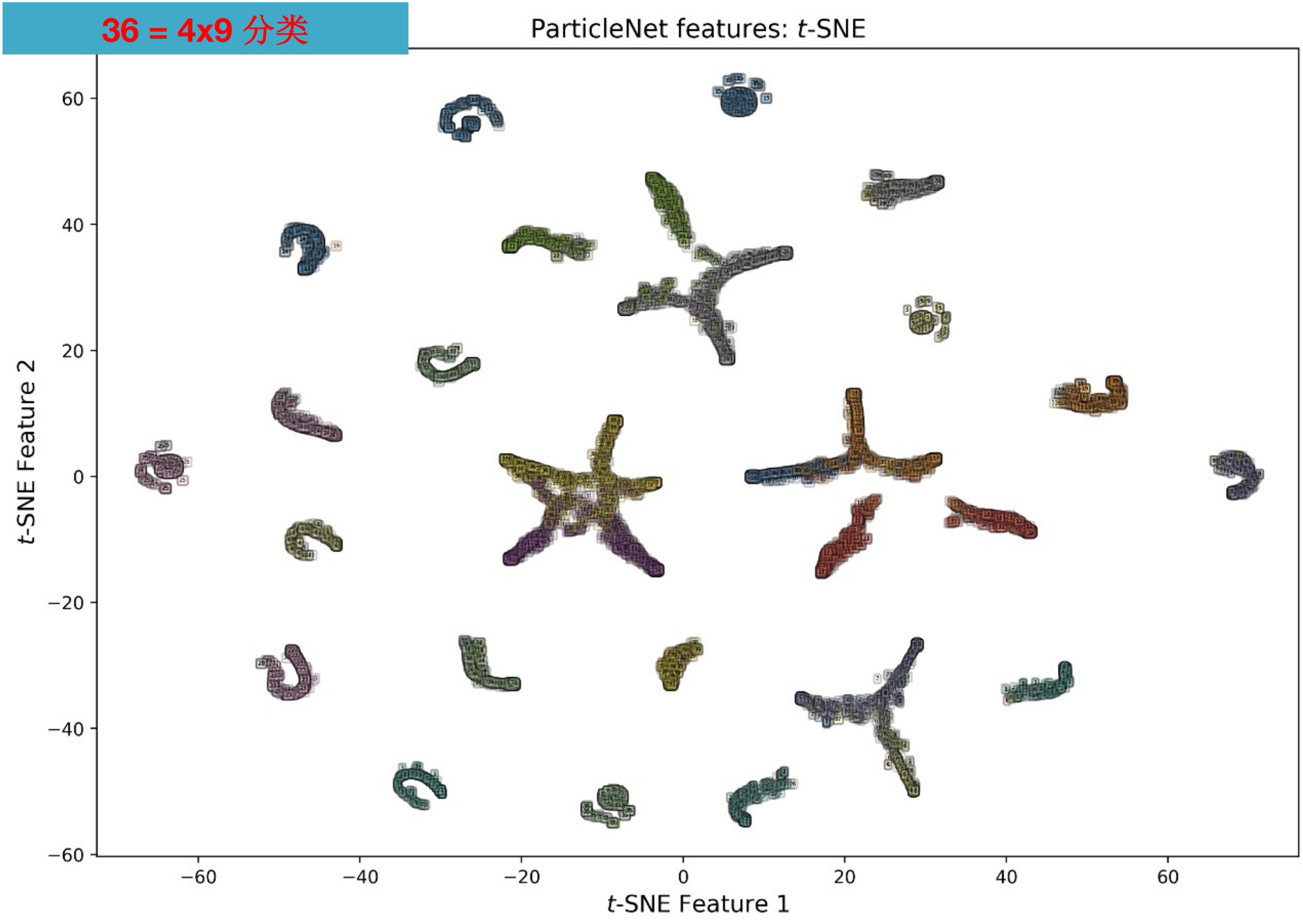

cc 0.86 0.05 0.00 0.00 0.02 0.00 0.02 0.04 0.01
bb 0.04 0.88 0.00 0.00 0.01 0.00 0.04 0.00 0.02
μμ 0.00 0.00 0.98 0.00 0.00 0.00 0.00 0.00 0.01
ττ 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.01
gg 0.04 0.02 0.00 0.00 0.82 0.00 0.05 0.07 0.01
γγ 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
ZZ 0.06 0.08 0.00 0.01 0.09 0.00 0.58 0.13 0.02
WW 0.05 0.00 0.00 0.01 0.08 0.00 0.08 0.77 0.01
γZ 0.01 0.01 0.01 0.02 0.01 0.00 0.02 0.01 0.91
True
cc bb μμ ττ gg γγ ZZ WW γZ
Predicted
0.0 0.2 0.4 0.6 0.8 1.0

(c)


cc 0.76 0.09 0.00 0.00 0.07 0.00 0.03 0.04 0.02
bb 0.11 0.73 0.00 0.00 0.09 0.00 0.05 0.01 0.02
μμ 0.00 0.00 0.98 0.00 0.00 0.00 0.00 0.00 0.01
ττ 0.00 0.00 0.00 0.98 0.00 0.00 0.01 0.00 0.01
gg 0.09 0.06 0.00 0.00 0.74 0.00 0.05 0.05 0.02
γγ 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00
ZZ 0.09 0.09 0.00 0.02 0.16 0.00 0.53 0.09 0.02
WW 0.08 0.04 0.00 0.01 0.13 0.00 0.09 0.63 0.01
γZ 0.02 0.02 0.01 0.02 0.02 0.00 0.02 0.01 0.90
True
cc bb μμ ττ gg γγ ZZ WW γZ
Predicted
0.0 0.2 0.4 0.6 0.8

(d)

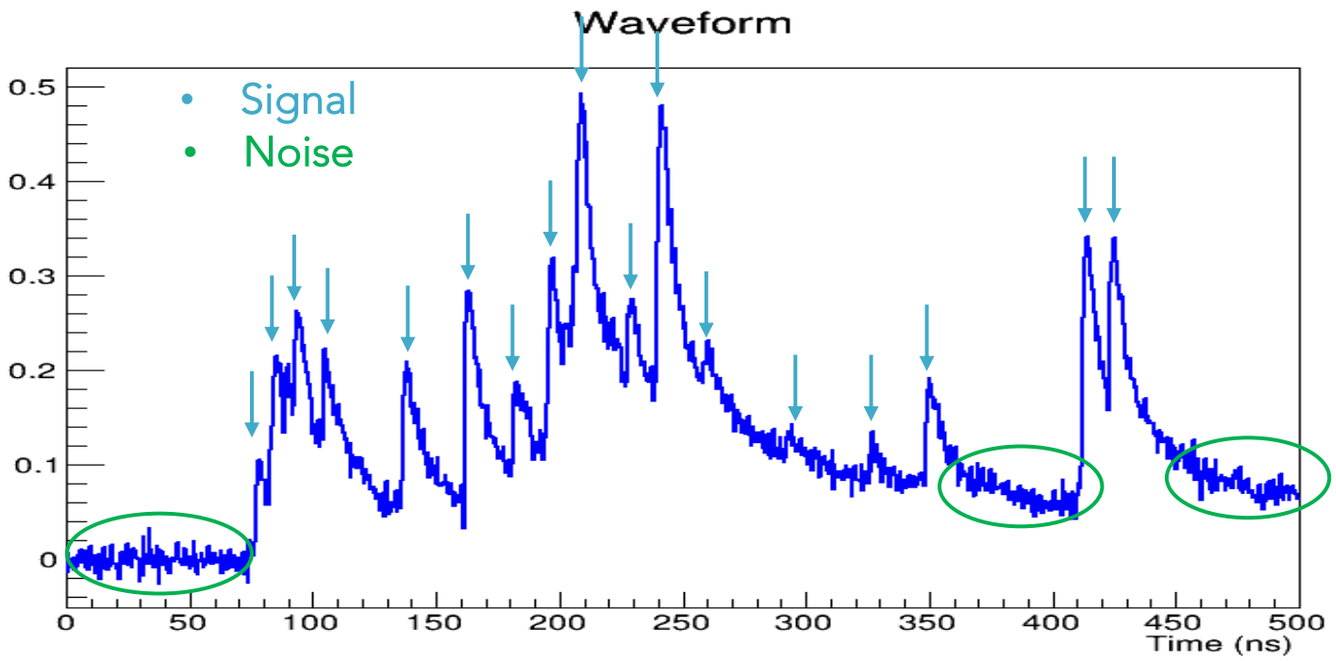
36 = 4x9 分类

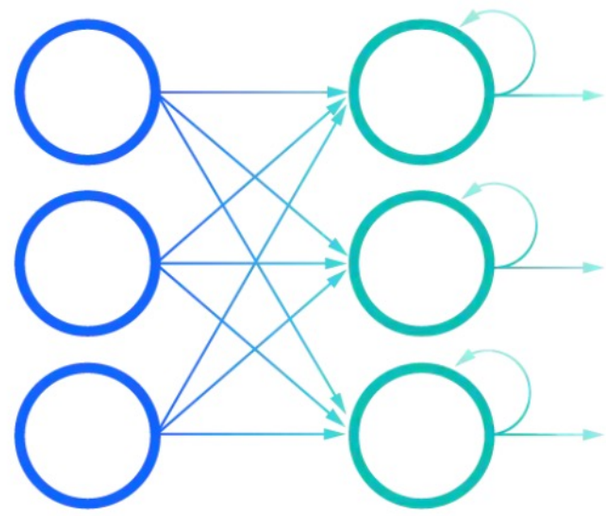

ParticleNet features: t-SNE
t-SNE Feature 2
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
t-SNE Feature 1

一个时序重建问题的例子

By 赵光， et al

- Peak detection of waveforms from the DC
- Supervised-classification: “signal” and “noise”





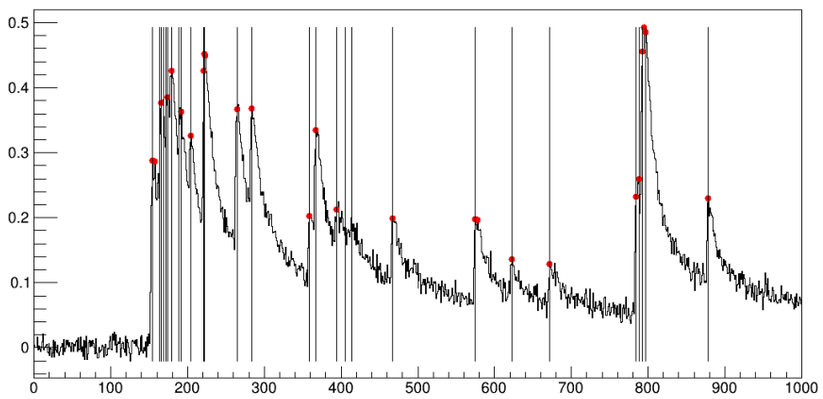
<u>Recurrent Neural Network (RNN)</u>:

- “Memory” structure: internal loops over sequence elements
- Powerful to handle time-sequences

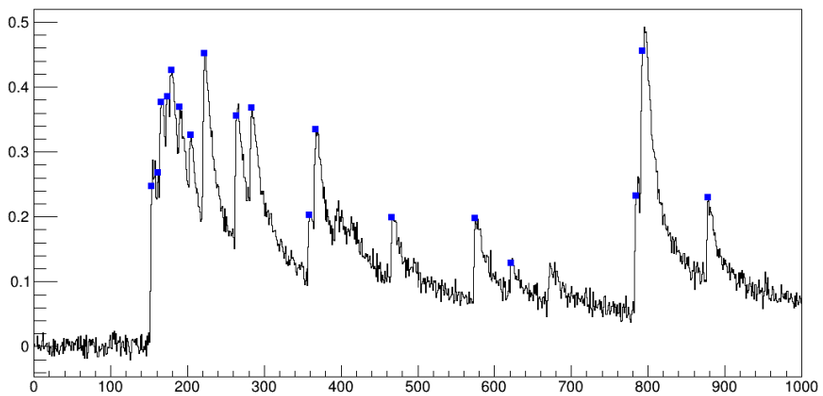
DL RESULTS AND COMPARE TO TRADITIONAL ALGORITHM

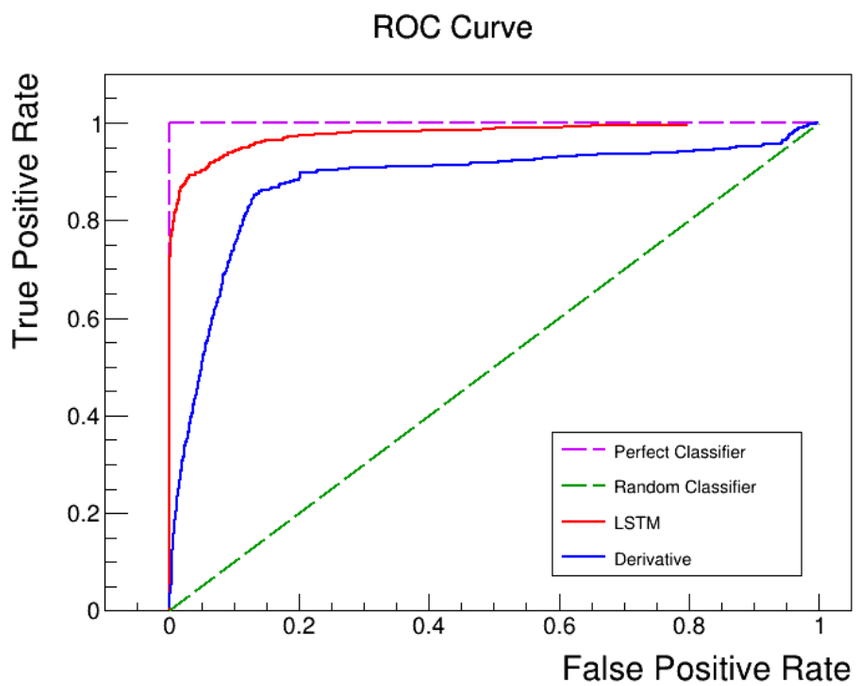
Black lines:
MC truth times

RNN (LSTM) 24/28 detected



Derivative 18/28 detected





RNN (LSTM) is much more powerful than the derivative for the peak finding problem

Thanks to Zhao Guang

Intelligent Readout of Pixel Sensors

By 卢云鹏

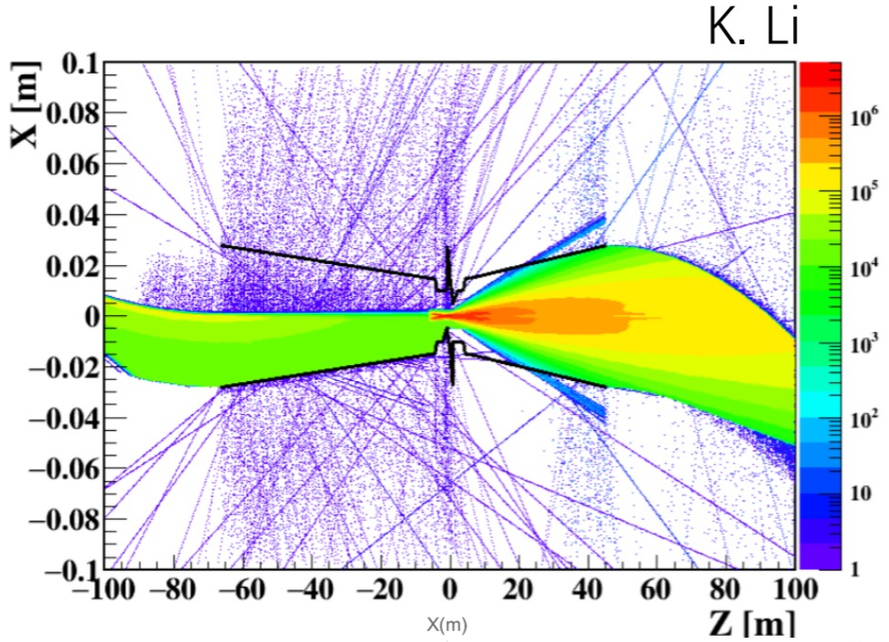
Challenges in the Vertex detector

- Data rate > Gbps / pixel chip, while power consumption limited < 50 mW/cm^2
  - 10 MHz particle hits / cm^2 at Z pole → 10 MHz * 3 pixels / cluster * 4 cm^2 / chip * 32 bit = 3.84 Gbps
  - High speed data link are always the hot spot of pixel chip
- The Neural Network was explored for possible solutions:
  - Data compression algorithm
  - Background suppression method

Background suppression

- Hit rate dominated by the radiative background for the CEPC vertex detector
- A pattern recognition module can be integrated into the pixel chip
  - Local hit pattern can be classified by a neural network
  - Algorithm developed with simulation data
  - Parameter reconfigurable based on the chip position and experimental data
- Data can be processed at hit level, a simple network is essential for low power operation

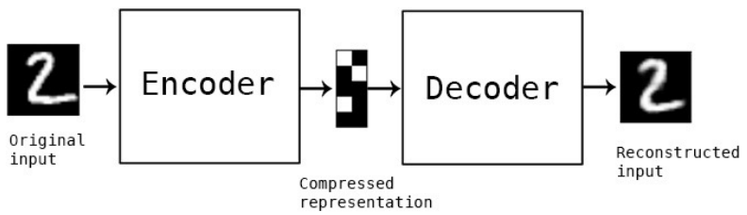
K. Li



Synchrotron radiation background

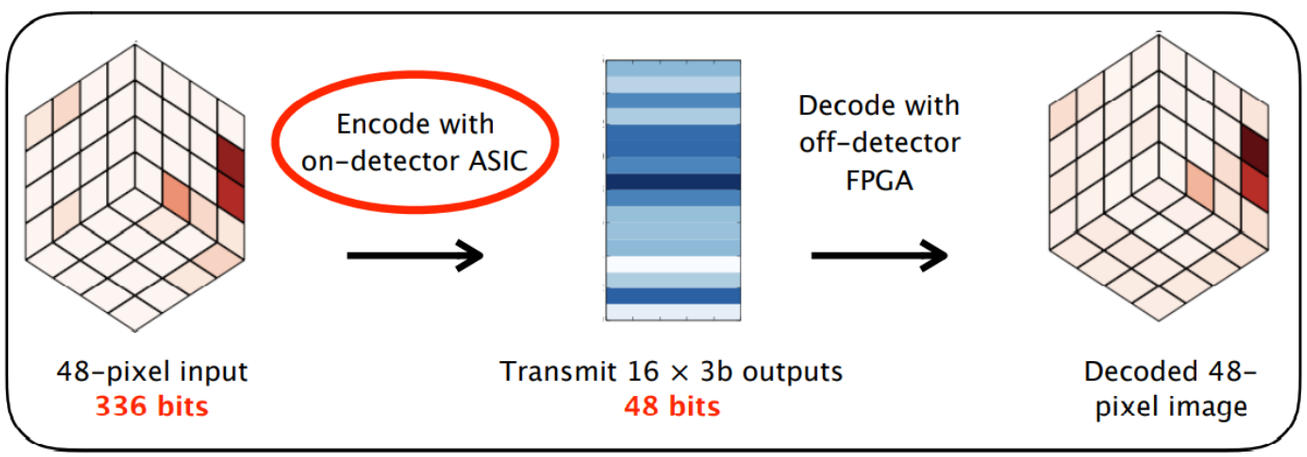
Autoencoder Neural Networks

- Compression algorithm, data-specific, lossy and **learned automatically**
  - https://blog.keras.io/building-autoencoders-in-keras.html
  - Being investigated by the High-Granularity Calorimeter Group
- Also considered for the data compression of CEPC vertex detector



  - Encoder on chip, and decoder in the back-end electronics or data processing software
  - Need to deal with much more channels and different data patterns

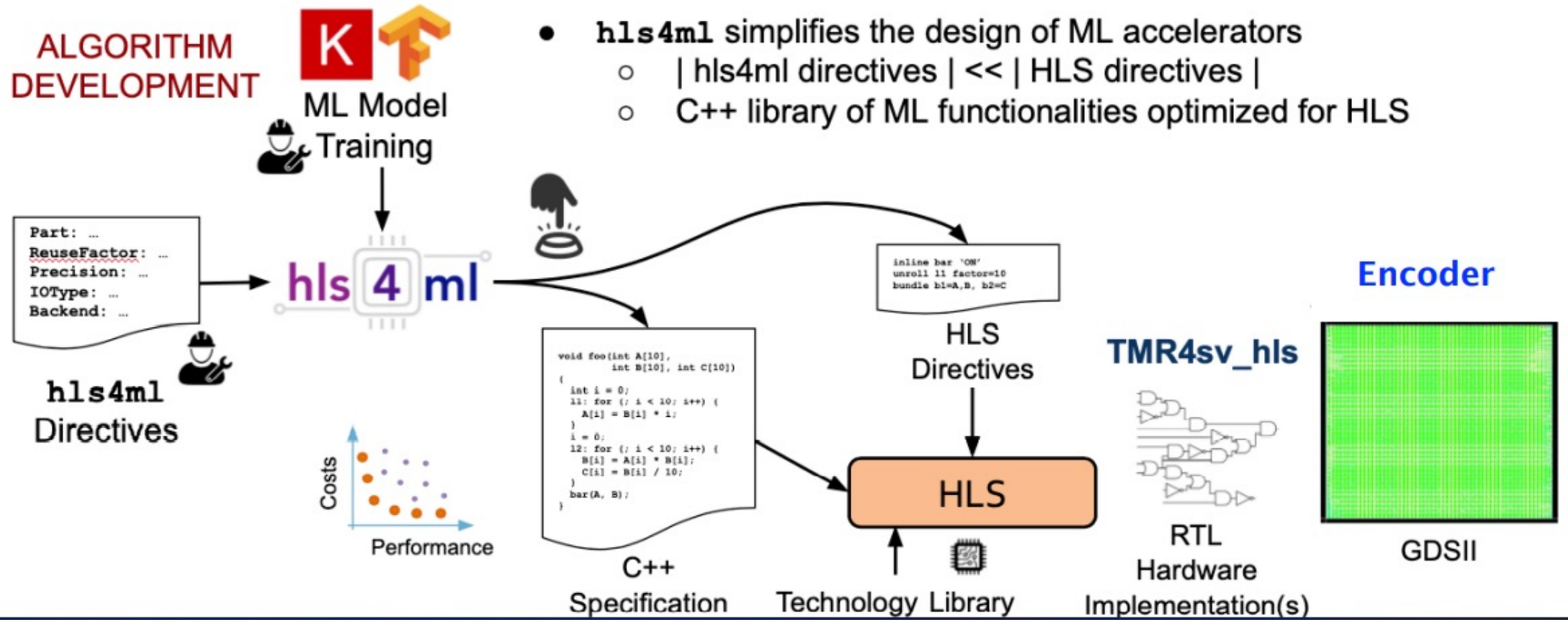
HGCAL 8" Module
Each trigger Cell consists of 3*3 sensors



Physics driven hardware co-design

Rapid prototyping and optimization of network achieved through

- **QKeras** : network development with **quantization-aware training** and physics simulation
- **hls4ml** : neural network description (h5 file e.g.) ➔ HLS-compliant C++ format
- **Catapult HLS** : C++ ➔ RTL
- **TMR4sv_hls** : Automated TMR for System Verilog



- **`hls4ml`** simplifies the design of ML accelerators
  - | hls4ml directives | << | HLS directives |
  - C++ library of ML functionalities optimized for HLS

Design of a reconfigurable autoencoder algorithm for detector front-end ASICs

Giuseppe Di Guglielmo

2020/11/30, Fast Machine Learning for Science Workshop

人力、经费

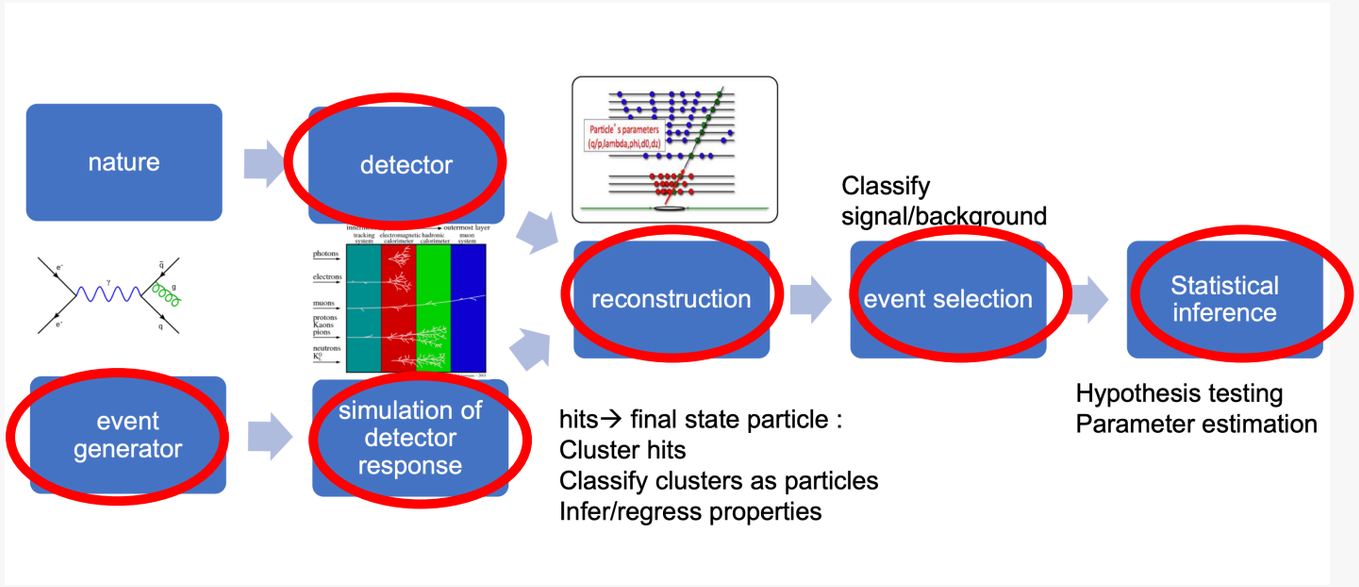
- 人员
  - 职工：赵光、方文兴、陆云鹏、李卫东、张瑶、李刚
  - 合作高校：东南大学（白羽）、吉林大学（宋维民）
- 一个面上：深度学习在 jet tagging方面的应用

计划

- 用 ML-aided E2E 分析实现 CEPC 探测器的快速优化迭代
- Jet energy resolution，jet charge
- Peaking finding
- Background suppression + data compression
- 自己的、更好的模型：快+好
- …

Summary

- 机器学习、深度学习很有用
- 机器学习不是黑盒子
- 需要更好理解物理问题和建模
- 需要学习机器学习相关知识：（高维）统计



- 可做的很多：分类，聚类，回归，异常探测，模拟，刻度，统计推断(likelihood-free)，...
- 可以让以前很多不可能的事情成为可能：多快好省
- 需要学习，需要和工业界交流
- 需要投入：人力，经费，...

用一个冷笑话结尾: ArXiv:1907.10621

- *Are you trying to replace PhD students with a machine?*
  As a preemptive safety measure against scientists being made redundant by automated inference algorithms, we have implemented a number of bugs in `MadMiner`. It will take skilled physicists to find them, ensuring safe jobs for a while. More seriously, just as `MadGraph` automated the process of generating events for an arbitrary hard scattering process, `MadMiner` aims to contribute to the automation of several steps in the inference chain. Both developments enhance the productivity of physicists.

Extras

推荐文献

- A high-bias, low-variance introduction to Machine Learning for physicists， Physics Reports 810 (2019) 1–124
- Relational inductive bias, deep learning, and graph networks, arXiv:1806.01261
- Geometric Deep Learning, 5Gs, arXiv:2104.13478